Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Scalable cell-specific coexpression networks for granular regulatory pattern discovery with NeighbourNet

Yidi Deng, Jiadong Mao, Jarny Choi, et al.

Genome Res. published online March 5, 2026
Access the most recent version at doi:10.1101/gr.281171.125

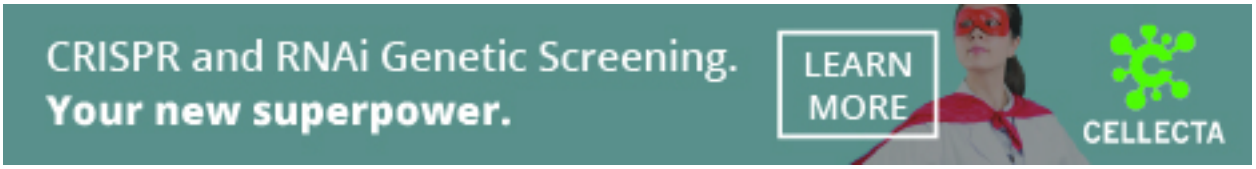
P<P Published online March 5, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Scalable cell-specific coexpression networks for granular regulatory pattern discovery with NeighbourNet

Yidi Deng[1,2], Jiadong Mao[1,†] & Jarny Choi[3,†] & Kim-Anh Lê Cao[1,*,†]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, 3010, Australia

[2] Research School of Finance, Actuarial Studies & Statistics, The Australian National University, 2601, Australia

[3]Bioinformatics and Cellular Genomics, St Vincent's Institute, 3065, Australia

† indicates equal contribution

∗ corresponding author: kimanh.lecao@unimelb.edu.au

Abstract

Gene networks provide a fundamental framework for understanding the molecular mechanisms that govern gene expression. Advances in single-cell RNA sequencing (scRNA-seq) have enabled network inference at cellular resolution; however, most existing approaches rely on predefined clusters or cell states, implicitly assuming static regulatory programs and potentially missing subtle, dynamic variation in regulation across individual cells. To address these limitations, we introduce NeighbourNet (NNet), a method that constructs cell-specific coexpression networks. NNet first applies principal component analysis to embed gene expression into a low-dimensional space, followed by local regression within each cell's k-nearest neighbourhood (KNN) to quantify coexpression. This approach improves computational efficiency and stabilises coexpression estimates, mitigating challenges posed by small sample sizes in KNN regression and the inherent noise and sparsity of scRNA-seq data. Beyond coexpression, NNet supports scalable downstream analyses, including (i) clustering and aggregating cell-specific networks into meta-networks that capture primary coexpression patterns, and (ii)

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

integrating prior knowledge to annotate coexpression and infer active signalling interactions at the individual cell level. All functional modules of NNet are implemented with an efficient algorithm that enables the application to large-scale single-cell datasets. We demonstrate NNet's effectiveness through three case studies on transcription factor activity prediction, early haematopoiesis, and tumour microenvironments. Provided as an R package, NNet offers a novel framework for exploring cellular variation in coexpression and integrates seamlessly with existing single-cell analysis workflows.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

Gene networks provide essential frameworks for understanding the complex molecular interactions that regulate gene expression in biological systems. At the core of these networks lies the inference of gene regulation, which reflects the binding of transcription factors (TFs) to specific DNA sequences to either activate or repress target gene expression. This regulation directs crucial cellular processes such as differentiation, proliferation, and responses to environmental stimuli. The emergence of high-throughput sequencing, particularly in transcriptomics in the early 2000s, has enabled rapid and affordable quantification of gene expression. Since then, substantial research efforts have been dedicated to developing statistical network inference methods that leverage associations in gene expression (coexpression) to decipher gene regulation (Huynh-Thu et al., 2010; Langfelder and Horvath, 2008). In such networks, edges represent measured coexpression between pairs of genes, and when these measurements carry regulatory (causal) implications, significant coexpression can be interpreted as evidence of gene regulation. These advances offer deep insights into cellular function and have contributed to the identification of key regulators of disease development, providing potential avenues for therapeutic intervention.

Single-cell RNA sequencing (scRNA-seq) has provided an unprecedented view of gene regulation by capturing gene expression profiles at the level of individual cells (Badia-i Mompel et al., 2023). Cells are often grouped into clusters based on their expression profiles to represent predefined cell states, and gene networks are inferred by measuring coexpression within these clusters to recover cell state-specific regulatory programs (Chan et al., 2017; Zhang et al., 2023; Morabito et al., 2023; Kamimoto et al., 2023). While clustering-based network inference methods have been the most widely investigated, they have a major limitation in assuming that cells within a cluster share similar and discrete regulatory programs. This assumption can mask subtle regulatory changes within each cluster and overlook fine-grained regulatory interactions that occur during cell state transitions. In addition, the accuracy of clustering-based methods relies heavily on the accuracy of cell clustering, meaning that poorly defined clusters can lead to biased or oversimplified findings.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

To overcome these limitations, researchers have developed cell-specific network (CSN) inference methods that measure coexpression within the local neighbourhood of each cell. CSN methods can be broadly classified into two categories based on how neighbouring cells are defined. Some methods define neighbouring cells based on their similarity in expression profiles (Zhang et al., 2022; Dai et al., 2019; Wang et al., 2021). Coexpression measures in these methods are typically based on correlation, meaning that they do not necessarily indicate gene regulation. Other methods define neighbouring cells along a developmental timeline, which can be inferred through pseudotime analysis (Zhang and Stumpf, 2023; Wang et al., 2023). In this approach, coexpression measures are time-dependent, potentially capturing causal (regulatory) relationships by associating gene expression at successive points in the trajectory. Despite their advantages, CSN inference faces key challenges in both scalability and stability. Inferring networks for thousands of cells is computationally intensive and often restricted to transcription factor interactions, limiting the biological scope of the networks. The intrinsic sparsity and noise of scRNA-seq data can also compromise the stability of coexpression estimates within small neighbourhoods. Moreover, many existing methods lack accessible implementations and offer only limited downstream analyses of inferred networks, hindering their practical application and reducing their impact on biological discovery (Pratapa et al., 2020; Nguyen et al., 2021).

To address these challenges, we present NeighbourNet (NNet), a novel method that uses k-nearest neighbours (KNN) principal component (PC) regression on gene expression data to efficiently construct robust, cell-specific coexpression networks. Intuitively, the coexpression between two genes in a given cell is quantified by how strongly one gene embedded within the PCs (i.e., predictor genes), contributes to the local regression model that predict the expression of the response gene within the cell's KNN. Our KNN-PC regression approach efficiently captures coexpression at scale. Rather than computing pairwise coexpression, it models a gene's expression using a small set of PCs, enabling simultaneous estimation of its coexpression with thousands of other genes, hence dramatically reducing computational cost. Embedding in PC space also mitigates the sparsity inherent in single-cell data, yielding more stable and noise-resistant coexpression estimates.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Together, these features allow NNet to construct large-scale, cell-specific networks that are both fast to compute and robust. Beyond network construction, NNet is the first framework to provide a comprehensive suite of downstream analyses on CSNs: meta-network analysis to uncover common coexpression patterns, meta-TF analysis to identify gene modules, and prior knowledge integration to facilitate cell-specific inference of active gene regulation and upstream signalling pathways (USPs). NNet, along with its downstream analysis modules, are available as an R package that integrates seamlessly with the widely used Seurat pipeline.

Results

Overview of NNet

We begin by outlining the workflow of NeighbourNet (NNet) and introducing the key terminology used in NNet analysis.

Coexpression network NNet estimates coexpression at the level of individual cells by combining dimensionality reduction with local regression. To measure cell-specific gene coexpression, NNet first uses principal component analysis (PCA) to embed gene expression data into a low-dimensional space (Figure 1A). Within this space, each cell finds its k-nearest neighbours (KNN), and a regression model is fitted to predict the expression of a response gene using the PCs. Within the neighbourhood of cell n, the coexpression level between the response gene p and a gene q contributing to PC computation (a predictor gene), denoted CSN_{npq}, is defined as how strongly the gene q influences the prediction of p through PCs given the fitted model.

By repeating this regression procedure across all response genes, NNet constructs a weighted coexpression network $\text{CSN}_{n..}$ for every individual cell, represented by a matrix of coexpression values between predictor and response genes for that cell (Figure 1B–C). Since NNet is regression-based, switching the roles of response and predictor genes yields different coexpression values for the same gene pair. The choice of responses depends on context, with the general rule of keeping the number of responses smaller than the number of predictors to reduce computational cost.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

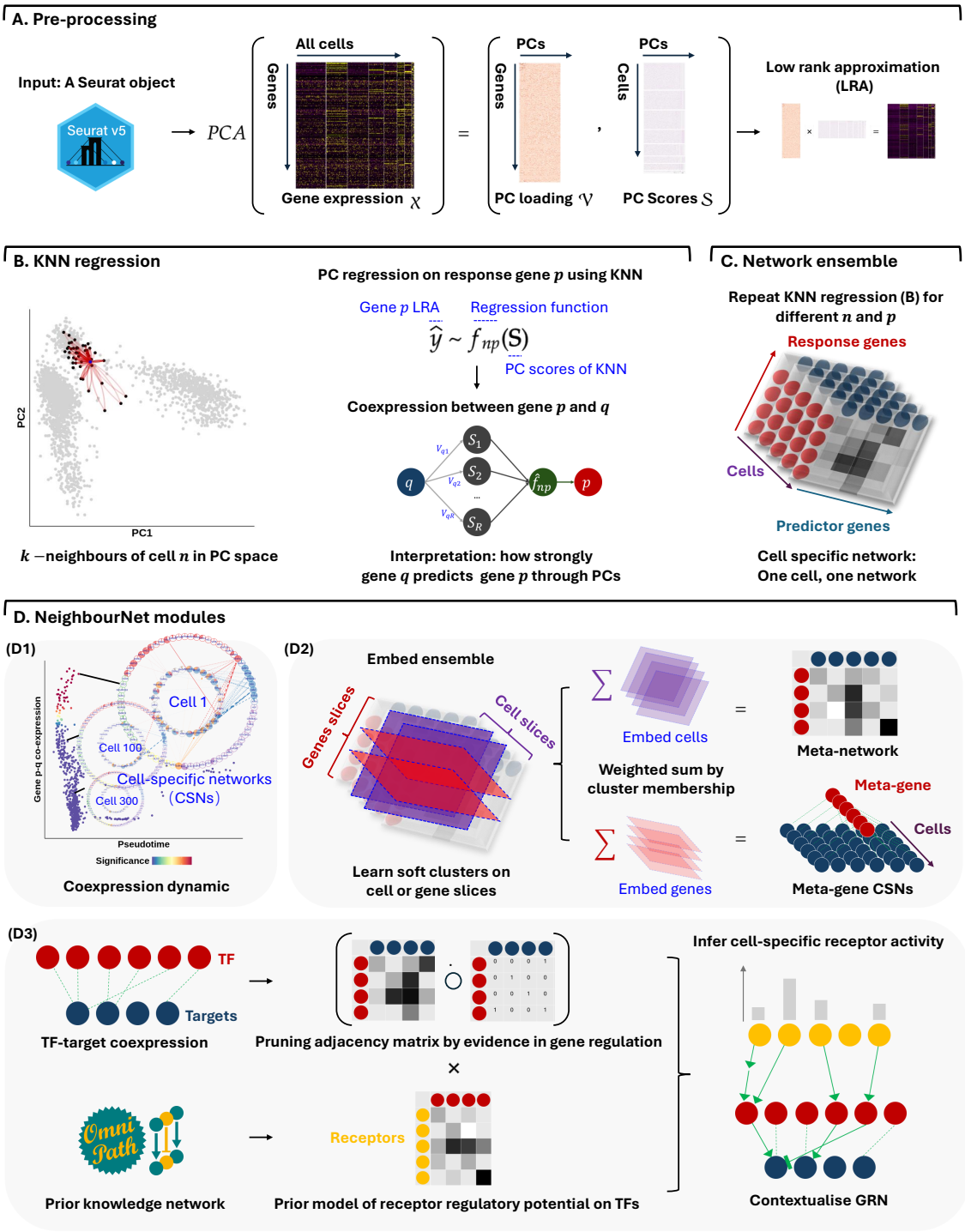


Figure 1. Caption next page.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 1. NeighbourNet (NNet) workflow for inferring cell-specific coexpression networks and integrating prior knowledge. A. preprocessing: Single-cell gene expression data are subject to principal component analysis (PCA) to capture the major variation. A low-rank approximation (LRA) then reconstructs the expression matrix using the PCs. A weighted k-nearest neighbour (KNN) graph that defines each cell's local neighbourhood in the PC space is constructed. **B. Neighbourhood Regression**: For each cell, NNet performs regression within the cell's KNN using the LRA-derived gene expression as the response and PCs as predictors. This process quantifies coexpression between genes by measuring how 'predictor genes' (those used to embed PCs) contribute to predict each response in the PC space. Repeating this for multiple response genes yields cell-specific coexpression networks (CSNs). **C. Network Ensemble Construction**: Collecting all CSNs form a network ensemble, which is a *cell* × *response* × *predictor* 3D array that can be diced and sliced. **D. Downstream Analysis**: (D1) We can slice the network ensemble to examine CSNs, analysing cellular variation and gene regulation dynamics through coexpression (D2) Non-negative matrix factorisation (NMF) identifies soft clusters of cells with similar network structures. Aggregating CSNs by these clusters yields meta-networks that summarise shared coexpression patterns. Same analysis strategy can be applied to the gene dimension corresponding to transcription factors (TFs), producing TF modules that co-regulate common programs. Aggregating TFs by module yields simplified meta-TF–target networks with improved interpretability. (D3) We adapted the NicheNet framework to integrate gene regulation and signalling interaction databases from OmniPath, constructing integrated prior knowledge networks (PKNs). Annotating CSNs with these PKNs transforms them into contextualised gene regulatory networks (GRNs). In addition, PKNs enable the inference of upstream signalling pathways (USPs) for each CSN by tracing signal transduction paths (receptor–TF–target) inferred based on the contextualised TF–target interactions.

Embedding network ensemble The collection of CSNs builds a network ensemble, which is a *cell*×*response*×*predictor* three-dimensional array (a 'cube') where each slice along the cell dimension corresponds to a CSN. While this per-cell resolution is powerful, it can also be helpful to extract higher-level structures that summarise thousands of such CSNs. To achieve this, NNet employs non-negative matrix factorisation (NMF) to identify coherent coexpression patterns across the ensemble, grouping cells that share similar network structures into soft clusters and summarising them as corresponding meta-networks (Figure 1D1,2).

Each soft cluster of cells is represented by a matrix $\mathcal{H}$, where each column $\mathcal{H}_{\cdot i}$ is a weighting vector defining cluster i, and each entry $\mathcal{H}_{ni}$ indicates the degree of membership of cell n to that cluster. NMF learns these soft clusters based on cell–cell similarity in their networks, and the corresponding meta-network for cluster i is computed as a weighted average of the individual CSNs:

$$\text{Meta-Network}_i = \sum_n \text{CSN}_{n\cdot\cdot}, \mathcal{H}_{ni}.$$

Our NMF approach extracts multiple soft clusters and meta-networks in sequential order, such that meta-network 1 captures the most significant coexpression pattern present in the dataset.

The same strategy can be applied to learn soft clusters of genes, summarising slices of the network ensemble along either the response or predictor dimension. When applied to TFs, each soft

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

cluster represents a TF module, which is a group of TFs that act together to regulate a common transcriptional program. Aggregating TFs according to their learned module yields meta-TF–target CSNs with only a few meta-TFs.

Signalling inference In parallel, we incorporate prior knowledge by mapping observed coexpression edges to curated regulatory and signalling interaction databases. This step produces context-specific gene regulatory networks (GRNs) with regulatory directories and allows for the inference of upstream signalling pathways (USPs) that connect ligand–receptor interactions to transcriptional responses through contextualised TF-target coexpression (Figure 1D3). The output includes interpretable networks, tools for visual exploration of networks' regulatory structures, and quantification of active receptor signalling across individual cells.

NNet coexpression between transcription factors (TFs) and targets provides robust evidence for detecting active gene regulation

NNet introduces a novel framework for measuring cell-specific coexpression (Figure 2A1). To evaluate the quality of the coexpression estimates obtained from PC regression and their relevance for gene regulation, we tested whether TF activity could be accurately inferred using NNet-derived TF–target coexpression profiles. While NNet does not calculate TF activity itself, these profiles can be used as input for established inference tools, namely AUCell and decoupleR (Aibar et al., 2017; Badia-i Mompel et al., 2022).

Traditional TF activity analysis, including AUCell and decoupleR, takes single-cell expression data as input and produces activity scores for each TF in each cell. For a given TF, the inference is based solely on the expression levels of its known targets: if a cell exhibits higher expression of those targets relative to other genes, the cell is assigned a higher activity score on that TF (Figure 2A2). However, high target expression does not necessarily indicate active TF regulation, often leading to false positives. Importantly, the same inference tools can also take cell-specific TF–target coexpression as input. By replacing target expression with coexpression profiles, NNet increases the robustness of TF activity inference: activity is now assessed by testing whether a TF is strongly

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

coexpressed with its known targets within each cell's neighbourhood, which is a signal more likely to reflect genuine regulatory activity (Figure 2A3).

We hypothesised that coexpression–based inference should carry stronger regulatory relevance than expression alone, and therefore produce TF activity scores that more accurately capture true cellular states. To validate this, we benchmarked activity scores derived from expression (ETAS, expression-based TF activity scores) against those derived from coexpression (CoETAS, coexpression-based TF activity scores).

Data and setting We analysed two datasets. The first was the PBMC 3K dataset from 10x Genomics, often used as a benchmark. The second dataset was a Perturb-seq experiment (Papalexi et al., 2021), in which the THP-1 monocyte cell line was subjected to pooled CRISPR screening to characterise the molecular networks regulating PD-L1 expression (Table 1).

For the PBMC dataset, our aim was to perform a sanity check to assess whether TF activity inference based on NNet-derived coexpression is biologically meaningful. Specifically, we analysed 288 TFs with sufficient target coverage according to the CollecTRI gene regulation database (Müller-Dott et al., 2023) (Material & Methods). Using NNet, we computed each TF's coexpression with 4,116 potential target genes (i.e., any gene with a known TF regulator in CollecTRI). From these coexpression profiles, we derived CoETAS and compared them with ETAS generated directly from target gene expression. We then evaluated which scores better preserved cell-type clustering and better highlighted known cell type–specific TFs. This comparison serves as a validation step, showing that NNet's coexpression measure provides sufficient biological signal to capture meaningful regulatory variation.

For the Perturb-seq dataset, which includes perturbations on 5 TFs across different subsets of cells, we again used NNet to measure coexpression between perturbed TFs and 5,151 target genes. Subsequently, we calculated ETAS and CoETAS for each of the 5 perturbed TFs across all cells. Because the dataset provides precise information on TF perturbations in each cell, it allowed us to directly assess how well the inferred activity scores distinguish perturbed from unperturbed cells. This, in turn, provides a hint of whether NNet's coexpression measure carries accurate information about TF regulatory status.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

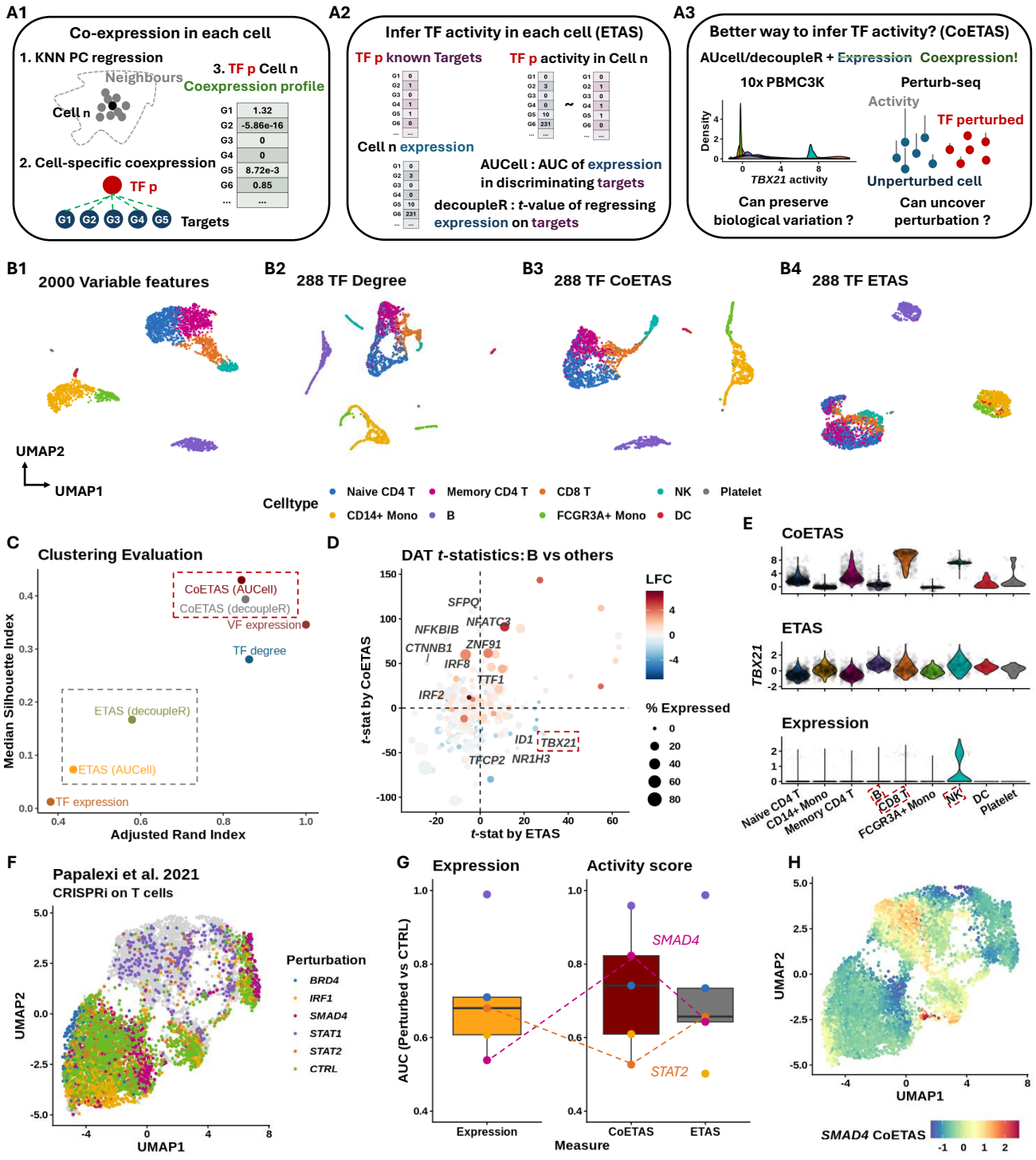


Figure 2. Caption next page.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2. Coexpression between transcriptional factors and targets provides robust evidence to active gene regulation. (A) (A1) NNet measures the coexpression between each transcription factor (TF) and its potential targets in each cell's neighbourhood, generating a TF–target coexpression profile. These profiles serve as input for TF activity inference. (A2) Traditional methods such as AUCell and decoupleR typically use target gene expression to infer TF activity, which can be misleading if high expression does not reflect active regulation. (A3) With NNet, AUCell or decoupleR are instead applied to coexpression profiles rather than expression profiles. This approach is expected to yield activity scores that better reflect true TF regulation. We benchmarked all approaches by comparing combinations of input (expression vs. coexpression) and method (AUCell vs. decoupleR). **(B)** UMAP visualisation of PBMC 3K data. Cells were embedded with (B1) the 2,000 most variable features (VFs), (B2) node degree for 288 TFs in cell-specific networks, (B3 vs. B4) TF activity inferred from coexpression (CoETAS) vs. expression (ETAS) using decoupleR. CoETAS preserves cell-type boundaries more effectively, resulting in clearer clustering. **(C)** The adjusted Rand and silhouette indexes were used to quantify clustering quality based on TF activity scores. Including the 2,000 VFs and 288 TF expression as controls confirms CoETAS's superior clustering performance, matching (B3). **(D)** *t*-tests on decoupleR-derived CoETAS and ETAS identified differentially activated TFs. The x- and y-axes show *t*-statistics from ETAS and CoETAS, respectively. TFs are coloured by their log-fold change expression in B cells vs. other cell types and sized by the percentage of B cells expressing the TF. Many TFs marked as active by ETAS alone (e.g., *TBX21*: a T/NK cell regulator) have low expression in B cells, suggesting false positives. **(E)** Comparing *TBX21* activity across cell types shows that ETAS (top) mistakenly marks *TBX21* as active in B cells, while CoETAS (middle) correctly highlights its known role in T/NK cells. The bottom panel shows *TBX21*'s log-normalised expression. **(F)** UMAP of the Perturb-seq dataset from Papalexi et al. (2021), with cells carrying TF perturbations are highlighted. **(G)** For 5 TF perturbations, ETAS and CoETAS were computed using decoupleR and evaluated by their ability to distinguish perturbed cells from controls using area under the curve (AUC). TF expression was served as a baseline. On average, CoETAS outperformed both ETAS and the baseline, with a particularly strong improvement for *SMAD4*. CoETAS performed poorly for *STAT2*, whose perturbed cells show little separation from controls in (F). **(H)** CoETAS scores for *SMAD4* show a distinct activity drop among *SMAD4*-perturbed cells.

NNet coexpression-based TF activity scores better preserves cell clusters in PBMC data. We annotated the dataset by first performing cell clustering based on the expression of the 2,000 most variable genes, followed by cell type assignment using cluster-specific marker genes (Figure 2B1). Using NNet, we constructed CSNs, which enabled novel analyses of cellular variation based on network topology, such as connectivity (i.e., node degree). A UMAP embedding of cells based on the connectivity profiles of 288 TFs within these networks largely preserved the separation of cell types (Figure 2B2).

Using both target gene expression and NNet TF-target coexpression, we computed TF activity scores and assessed how well these scores preserve cell type clusters. UMAP embeddings showed that CoETAS (Figure 2B3) better captured known cell types compared to ETAS (Figure 2B4, Supplementary Figure S1A). Quantitative evaluations using the adjusted Rand index and the silhouette index further confirmed that CoETAS produced higher-quality clustering (Figure 2C). As a control, clustering based on TF expression performed worse than TF activity. Additional evaluations using clusters derived from low-rank approximations of TF expression (Supplementary

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure S1B) ruled out the possibility that the advantage of CoETAS stemmed from the PCA regression used to measure coexpression. These results highlight the ability of NNet coexpression in capturing cell-specific regulatory activity.

NNet coexpression detects more meaningful cell-type-specific gene regulation in PBMC data. We then assessed the biological relevance of differentially activated TFs (DATs) identified by activity scores in different cell type clusters. We focused on B cells, which formed the most distinct cluster within the PBMC data (Figure 2B). We conducted *t*-tests on activity scores to extract DATs in B cells. The *t*-tests on ETAS and CoETAS revealed largely different sets of DATs (Figure 2D). While most DATs identified by both methods were differentially expressed by B cells, several DATs identified solely by ETAS were actually downregulated or not expressed at all by the B cell population. A key example was *TBX21*, a master regulator of T and NK cells (Szabo et al., 2000; Knox et al., 2019), which was deemed active in B cells by ETAS. In contrast, CoETAS correctly highlighted the relevance of *TBX21* in T cells and NK cells (Figure 2E). Thus, relying solely on target expression for TF activity inference can yield spurious results, whereas NNet coexpression provides more accurate evidence for genuine regulatory activity.

NNet coexpression based TF activity scores distinguish perturbation effects more effectively in the Perturb-seq data. We further assessed the robustness of CoETAS by analysing TF perturbation (knockdown) experiments in the THP-1 cell line. A UMAP embedding of cells showed variable responses to five TF perturbations, with *STAT1* knockdown producing the strongest effect, as perturbed cells were clearly separated from the control cluster (Figure 2F). For each perturbed TF, we computed CoETAS and ETAS using decoupleR, and used area under the curve (AUC) to evaluate their ability to distinguish perturbed cells (expected lower activity; negative class) from controls (expected higher activity; positive class). On average, CoETAS outperformed both ETAS and raw TF expression, achieving higher AUC scores (Figure 2G). The largest performance gain was observed for *SMAD4*, where CoETAS clearly identified reduced activity in *SMAD4*-perturbed cells (Figure 2H), whereas ETAS patterns were less distinct (Supplementary Figure S1C1). In contrast, CoETAS performed worst for *STAT2*, which is likely due to the weak

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

perturbation effect suggested by the close proximity of *STAT2*-perturbed cells to controls in the UMAP. Nevertheless, CoETAS still highlighted a subcluster enriched for *STAT2*-perturbed cells (Supplementary Figure S1C2).

We performed a similar analysis on an independent Perturb-seq dataset (Dixit et al., 2016) containing more TF perturbations (Supplementary Results; Supplementary Figure S1) and confirmed the advantage of CoETAS over ETAS in detecting perturbation effects. .

Overall, our findings show that incorporating NNet coexpression significantly enhances the accuracy and biological relevance of TF activity inference, underscoring its potential in effectively capturing gene regulation activities. Additional analysis comparing NNet with two other cell-specific methods, oCSN (Dai et al., 2019) and LocCSN (Wang et al., 2021), revealed that NNet substantially improves CSN construction by better preserving cell-type identity (Supplementary Results; Supplementary Figure S7).

NNet cell-specific coexpression networks enable granular profiling of dynamic alternation in gene regulation.

Using a stem cell biology case study, we demonstrate that NNet resolves cellular variation at the level of cell-specific coexpression (Figure 3A1). We propose an efficient matrix factorisation technique to summarise and interpret large-scale coexpression data as meta-networks and meta-genes, revealing overarching coexpression patterns and principal regulatory programs (Figure 3A2). We demonstrate that the NNet framework facilitates accurate identification of pivotal points in cell fate decisions and enhances our understanding of the underlying transcriptional mechanisms that coordinate these decisions.

Data and setting We used the lineage-negative (Lin^-) bone marrow cell scRNA-seq data from (Pellin et al., 2019), which was designed for profiling the early human haematopoietic landscape (Table 1). With TF and target gene sets acquired from our integrated prior knowledge network (PKN) of gene regulation (Material & Methods), NNet analysis was performed to measure coexpression

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

between 805 TFs and 4,600 potential targets.

In the following sections, we perform meta-network analysis and meta-TF analysis separately to study two distinct major lineages: the erythroid lineage and the mononuclear phagocyte (MP) lineage. Meta-network analysis is used to examine transcriptional program alterations during erythropoiesis, while meta-TF analysis is applied to identify the key TF module that potentially coordinates MP differentiation.

NNet meta-network analysis reveals abrupt rewiring of coexpression network during erythrocyte maturation. We identified 20 meta-networks, each characterised by a set of cell weights indicating the cell populations (i.e., soft cluster) they represent (Supplementary Figure S2A; see Material & Methods for a detailed definition of cell weights). These meta-networks captured key transcriptional shifts during hematopoiesis. Specifically, the first four meta-networks delineated the primary myelopoietic lineages towards erythroid (Figure 3B) and MP fates (Supplementary Figure S2B,C). For example, meta-network 1, primarily composed of networks of megakaryocyte–erythroid progenitors (MEPs), was associated with early lineage commitment characterised by *HDAC7*-driven suppression of *MYC* that releases erythroid differentiation blockage (Figure 3C) (Jayapal et al., 2010; Delgado and León, 2010; Wang et al., 2020). In contrast, meta-network 2, which exhibited increased involvement during terminal erythropoiesis, highlighted a shift towards network modules related to haemoglobin synthesis (e.g., *KLF1*-mediated *ABCB10* expression) (Tallack et al., 2010) and erythrocyte membrane protein synthesis (e.g. *TAL1*-mediated *RHD* and *ERMAP* expression) (Kassouf et al., 2010). Furthermore, the cell weights on meta-network 2 sharply localised to erythroid progenitors (ErPs) from MEPs, suggesting that erythrocyte maturation may involve abrupt, temporally distinct transcriptional reprogramming. Our results recapitulate a well-established TF regulation mechanism known as the 'GATA switch', in which the repression of *GATA2* facilitates a surge in *GATA1* expression: a crucial step in erythrocyte maturation (Bresnick et al., 2018; Suzuki et al., 2013).

NNet network topology of individual cells better characterises temporal dynamics of GATA regulation compared to gene expression. Focusing on the GATA genes (*GATA1* and

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

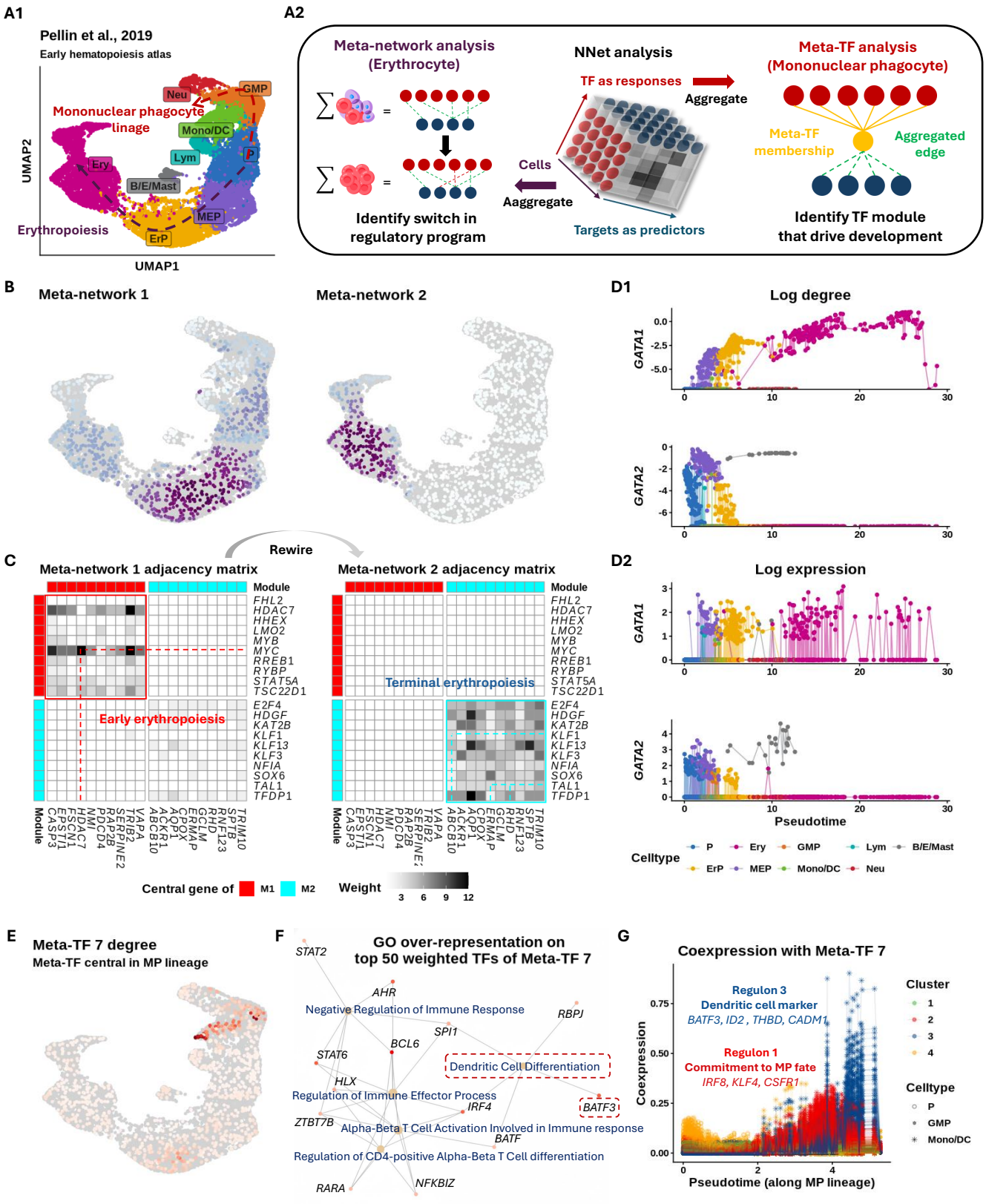


Figure 3. Caption next page.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 3. Dynamic coexpression network reveal key transcriptional program alternation during early hematopoiesis. (A) UMAP visualisation of bone marrow Lin^- cells, depicting the transcriptional landscape of early hematopoiesis (Pellin et al., 2019). P: progenitor, GMP: granulocyte-monocyte progenitor, MP: mononuclear phagocyte, Neu: neutrophil, Lym: lymphoid cell, Mast: Mast cell, MEP: megakaryocyte–erythroid progenitor, ErP: erythroid progenitor, Ery: erythrocyte. **(B)** Bipartite coexpression networks were constructed for individual cells, linking transcription factors (TFs) to target genes. Non-negative matrix factorisation (NMF) was applied to (i) embed the cell dimension into meta-networks and (ii) or to embed the TF dimension into meta-TFs. Meta-network and meta-TF analyses were used to resolve erythropoiesis (B–C) and mononuclear phagocyte (MP) differentiation (E–G). **(B)** Meta-network analysis assigns cell weights that define soft clusters of cells based on their coexpression patterns. Projecting the soft clusters onto a UMAP shows that the first two meta-networks capture major transcriptional program shifts during erythropoiesis. **(C)** Aggregating CSNs according to the soft clusters in (B) yields meta-networks 1/2 (early/late erythropoiesis) representing the clusters. Weighted adjacency matrices of meta-networks are shown as heatmaps. Top 10 TFs (rows) and targets (columns) from each meta-network reveal rewiring from a stemness-associated state (meta-network 1) to erythrocyte-autonomous programs (meta-network 2), indicating an abrupt transcriptional transition. **(D)** Continuous tracking of GATA gene (GATA1 and GATA2) regulatory dynamics during erythropoiesis. (D1) tracks changes in connectivity (log node degree) for GATA genes within CSNs during erythropoiesis. Compared to the gene expression patterns shown in (D2), connectivity provides a more precise representation of the regulatory dynamics of these TFs, highlighting a distinct switching behaviour between GATA2 and GATA1. **(E)** Meta-TF analysis assigns TF weights that define modules (i.e., soft clusters) of TFs based on their coexpression patterns. Aggregating TFs according to these modules yields simplified CSNs that summarise meta-TF–target coexpression. Degree of meta-TF 7 across cells were projected onto the UMAP. Meta-TF 7 emerged as the first module associated with the MP lineage, displaying the highest connectivity within GMPs. **(F)** Over-representation analysis of the top 50 TFs that composes meta-TF 7 identifies immune differentiation pathways. TFs linked to the enriched terms are coloured based on their weights on meta-TF 7. **(G)** Meta-TF 7's coexpression with each target gene along pseudotime. Each scatter path represents a target. Targets were clustered into four groups based on their coexpression patterns, with clusters 1 and 3 showing staggered activation waves linked to MP commitment and $CD141^+$ dendritic cell maturation.

GATA2), we explore how NNet network topology analysis provides an enhanced characterisation of dynamics of transcriptional program compared to conventional gene expression analysis. We calculated the connectivity of GATA genes within CSNs and plotted this connectivity against the diffusion pseudotime of cells (Figure 3D1) (Haghverdi et al., 2016). This analysis showed that *GATA2* reached its peak connectivity at the MEP stage, which then sharply declined following the transition to the ErP stage. As *GATA2*'s connectivity dropped to zero, *GATA1* began to show a significant increase in connectivity throughout maturation, emphasising a critical temporal coordination of *GATA2* and *GATA1* during erythropoiesis. In contrast, plotting the expression of GATA genes against pseudotime lacks critical details about the dynamics of GATA regulation, such as the peak of activity and the switching point between transcriptional states (Figure 3D2).

NNet Meta-TF analysis identifies key transcriptional programs that potentially direct $CD141^+$ dendritic cell differentiation. We proceed to illustrate the power of meta-TF analysis, focusing on the identification of key meta-TFs and their regulons associated with MP differentiation,

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

a process distinct from erythropoiesis. Each meta-TF represents a TF module (i.e., a soft cluster) defined by a set of TF weights. While analysing a single TF's coexpression often provides only a partial view of its function, interpreting TFs within the context of their modules offers integrative view of their role as a regulatory program.

In each CSN, TFs are aggregated according to modules to analyse meta-TF level coexpression with targets (Supplementary Figure S2D). By examining the connectivity of meta-TFs within each cell, we identified Meta-TF 7 as the first one that showed strong connectivity across the MP lineage, particularly in granulocyte–monocyte progenitors (GMPs) (Figure 3E; Supplementary Figure S2E). This suggests a potential role for Meta-TF 7 in MP differentiation. To functionally characterise meta-TF 7, we performed a Gene Ontology (GO) over-representation analysis on its module composition. This analysis revealed that meta-TF 7 was closely associated with immune cell differentiation (Figure 3F), particularly the differentiation of dendritic cells (DCs), which belong to the MP lineage. Notably, the DC differentiation term was enriched by *BATF3*, a key TF that is essential for the development of conventional type 1 DCs (cDC1) (Hildner et al., 2008; Satpathy et al., 2012). This clinically significant DC subtype plays a critical role in anti-cancer immunity (Böttcher et al., 2018), with extensive research dedicated to developing protocols for its in vitro differentiation (Rosa et al., 2022; Elahi et al., 2022). We then sought to identify and comprehend potential downstream targets of meta-TF 7 to gain a deeper understanding of its function. Figure 3G illustrates the changes in target gene coexpression with meta-TF 7 during MP differentiation. By clustering targets based on these coexpression values, we identified two major regulons that exhibited staggered waves of activity (Supplementary Figure S2F).

The first wave of activity, represented by regulon 1, included *IRF8*, *KLF4*, and *CSF1R*, which are critical markers indicative of MP fate commitment (Combes et al., 2021; Kurotaki et al., 2014; Tussiwand and Gautier, 2015). The second wave, represented by regulon 4, featured *BATF3* alongside other key markers of cDC1 identity, including *ID2*, a TF that delineates the cDC1 lineage (Jackson et al., 2011); *THBD*, which encodes the CD141 surface marker (Anastasiou et al., 2012); and *CADM1*, which is expressed exclusively by cDC1 (Collin and Bigley, 2018). These findings further support the role of meta-TF 7 in coordinating the transcriptional programs driving cDC1

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

differentiation. Notably, our analysis of marker expression revealed that cDC1 constituted only a small population at the tip of the MP lineage, making it challenging to identify through clustering analysis alone. This underscores the effectiveness of our meta-TF analysis in disentangling relevant TF-target interactions within rare cell populations.

In conclusion, our analysis of the early hematopoiesis atlas demonstrates NNet's capability to analyse cellular variation through coexpression, particularly providing insights into the dynamics of gene regulation throughout cell differentiation. By meta-network (meta-TF) analysis, we successfully identified principal coexpression patterns, TF modules and regulons at the individual cell level without the need for cell clustering.

NNet effectively integrates prior knowledge to facilitate cell-specific inference of upstream gene regulation signals

Coexpression between genes does not necessarily imply causal relationships in gene regulation. To enhance the interpretability of coexpression networks, we extend NNet to annotate TF-target coexpression by leveraging prior knowledge of gene regulation and signalling interactions. This approach helps us discern active gene regulation from confounding effects. In addition to identifying regulatory relationships, we infer upstream signalling pathways (USPs), the intracellular signal transduction events that start from receptors and lead to the TF-mediated gene regulation. Moreover, since NNet generates CSNs, the NNet USP inference can also be conducted at the individual cell level, allowing us to further infer signalling dynamics that contribute to cell state transitions.

For our third case study, we analyse scRNA-seq data of small cell lung cancer (SCLC), with a specific focus on its macrophage subset (Figure 4A1). We demonstrate how NNet with prior knowledge annotation comprehensively reveals critical signal interactions within the tumour microenvironment (TME), which may contribute to the development of tumour-associated macrophage (TAM) phenotypes (Figure 4A2).

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

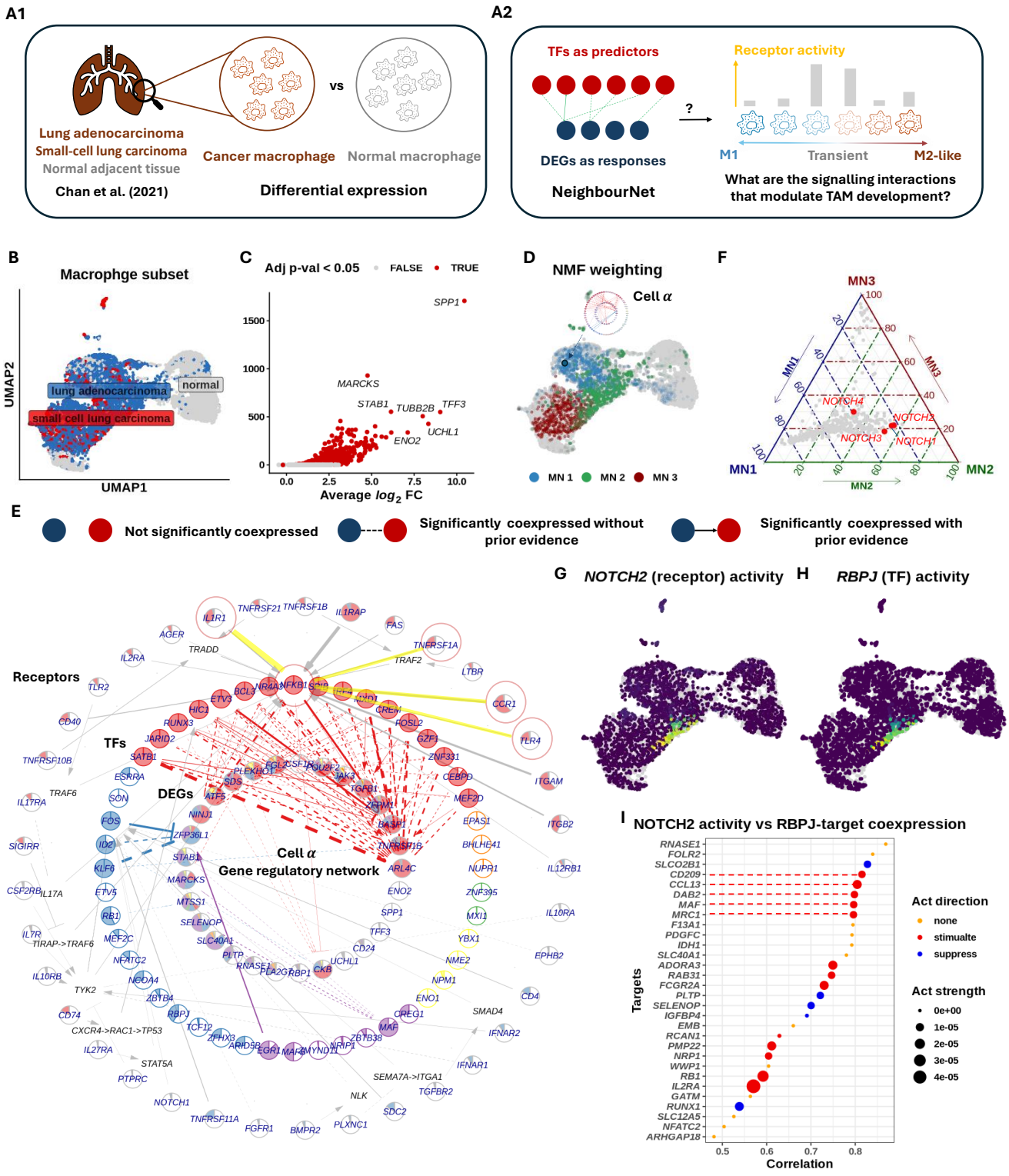


Figure 4. Caption next page.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 4. Prior knowledge annotation highlights critical signalling interactions facilitate tumour associated macrophage development. (A)(A1) We focused on the macrophage subset of Chan et al. (2021)'s small cell lung cancer scRNA-seq dataset. Using NNet, we aimed to identify the signalling interactions that regulate differentially expressed genes (DEGs) between cancer and normal macrophages, providing insights into how inter- and intracellular signalling interactions shape macrophage identity within the tumour microenvironment. (A2) We utilised NNet's upstream signalling pathway (USP) inference to identify receptors with high activity in driving transcription factors (TF) regulation of DEGs in transient macrophage populations during TAM development. **(B)** UMAP visualisation of macrophages in the dataset. **(C)** Volcano plot of the DE result. Only genes that were upregulated in cancer are shown. Pi value: negative $\log_{10}$ p-value times $\log_2$ fold change. **(D)** Soft clusters of cells for the first three meta-networks, illustrating a key transition from pro-inflammatory macrophages to TAMs. Cell α, with the highest cell weight on meta-network 1, is highlighted. **(E)** Annotated coexpression network of cell α. The innermost layer contains response (target) genes, encircled by central TFs in the meta-networks. Receptors occupy the outermost layer, each linked to a TF inferred to mediate that receptor's influence. If the receptor–TF link is indirect, an extra layer shows the shortest signalling path according to prior knowledge. Material & Methods provides details. This network captures the canonical NF-κB pathway in pro-inflammatory macrophages. **(F)** Ternary plot of receptor activity scores on the first three meta-networks. NOTCH receptors (*NOTCH2/3/4*) show peak activity in meta-network 2. **(G)** *NOTCH2* activity and **(H)** *RBPJ* connectivity projected onto the UMAP. *RBPJ* was the TF whose connectivity was found mostly correlated with *NOTCH2* activity. **(I)** Correlation between *NOTCH2* activity and different targets' coexpression with *RBPJ*. The top targets with the highest correlations are illustrated. *NOTCH2* activity is strongly associated with M2-like TAM marker expression mediated by *RBPJ*.

Data and setting. We acquired the SCLC atlas from Chan et al. (2021), which includes scRNA-seq data of both tumour and tumour-adjacent normal tissues (Supplementary Figure S3A; Table 1). Focusing on the macrophage subset (Figure 4B), we conducted a differential expression (DE) analysis comparing macrophages from tumour and normal samples. Among the top 50 genes upregulated by tumour macrophages, we identified 29 genes that are targets of known TFs according to our PKN of gene regulation (Figure 4C). The coexpression of these 29 genes with 900 TFs obtained from the PKN was then measured using NNet. We applied meta-network analysis to recover the major coexpression patterns and subsequently utilised USP inference to investigate how signals are transmitted to the 29 genes through TFs across different macrophage landscapes.

NNet annotated coexpression networks capture key signalling pathways defining macrophage interactions within TME. We focused on the top three meta-networks and projected the soft clusters they represent onto a UMAP embedding of macrophages (Figure 4D). These soft clusters distinctly localised to different macrophage populations, suggesting that each population is governed by a unique transcriptional program. To further explore signalling differences among these populations, we visualised the annotated coexpression networks of the three most representative cells (those with the highest cell weight on each meta-network), highlighting a spectrum of macrophage identities (Supplementary Figure S3B).

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

On one end of the spectrum, cell α, representing meta-network 1, showed a high level of connection across diverse regulatory modules (Figure 4E). The USP analysis revealed an enrichment of pro-inflammatory signalling in cell α, with *NFKB1* acting as the central TF that mediated signals from receptors like IL1R (*IL1R1*), TNFR1 (*TNFRSF1A*), and TLR4 (*TLR4*). This observation suggests the activation of the well-known NF-κB axis, which strongly supports the M1 pro-inflammatory identity (as opposed to the M2 anti-inflammatory identity) of cell α (Yu et al., 2020; Hagemann et al., 2009).

On the other end of the spectrum, cell γ, representing meta-network 3, displayed a complete loss of pro-inflammatory signalling (Supplementary Figure S3C1). Instead, it established *ENO1*-mediated signalling to stimulate *SPP1* expression, with SRB1 (*SCARB1*) as the primary receptor and *SRC* likely acting as the transducer relaying the signal to *ENO1*. *SPP1*$^+$ macrophages represent a distinct subtype of TAM that has been reported to interact with fibroblasts and anti-inflammatory CD8$^+$ T cells, driving tumour growth, metastasis, and immunosuppression (Qi et al., 2022; Bill et al., 2023). While the roles of SRB1, *SRC*, and *ENO1* in TAM have been independently reported, their interaction in promoting *SPP1* expression presents a potential regulatory mechanism that warrants further investigation (Plebanek et al., 2018; Dwyer et al., 2017; Liang et al., 2023).

In between cell α and γ, cell β, representing meta-network 2, exhibited an intermediate transcriptional state with an increase in the activity of the IL4R (*IL4R*) receptor, which is a critical switch for M2 state activation (Supplementary Figure S3C2). We hypothesise that cell β and the macrophage population associated with meta-network 2 exist in a highly plastic state, rendering them susceptible to reprogramming and valuable targets for therapeutic treatment. To better understand this transitional state, our next analysis investigates the key signalling interactions and their consequences among these transient macrophages in the TME.

NNet receptor activity analysis highlights activation of *RBPJ*-mediated NOTCH signalling during transitional macrophage state. NNet's USP inference involves calculating each receptor's potential activity across target genes for each cell, summarised as activity scores. Leveraging this inference, we explored the receptors that exhibit differential signalling in transient macrophages, represented by cell β, compared to M1 macrophages (e.g. cell α) and the TAM

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

population (e.g. cell γ). Specifically, we summarised receptor activity within each macrophage population by a weighted sum (using importance scores) of cell-specific receptor activities (Material & Methods). Our comparison revealed that NOTCH (*NOTCH2/3/4*) receptors were exclusively active in the transient macrophage population associated with meta-network 2 (Figure 4F). Endothelial cells, expressing the NOTCH ligands *DLL* and *JAG*, emerged as the principal source of NOTCH signalling (Supplementary Figure S3D). This finding aligns with previous research showing that NOTCH signalling can influence macrophage polarisation towards the M1 or M2 phenotype, although its precise direction remains a debated topic (Palaga et al., 2018; Chen et al., 2023; Xu et al., 2012; Foldi et al., 2016).

To clarify the role of NOTCH in our lung cancer setting, we examined TF connectivity in relation to *NOTCH2* activity (Figure 4G). *RBPJ* emerged as the top TF whose connectivity highly correlated with *NOTCH2* activity (Figure 4H), which is consistent with its established role as a key mediator of NOTCH signalling and a frequent target for blocking the pathway (Friedrich et al., 2022). Our findings partially recapitulate those of Franklin et al. (2014), who reported that *RBPJ*-mediated NOTCH signalling is crucial for the final differentiation of TAMs from tumour-infiltrating monocytes. However, it remains unclear whether this pathway also governs the macrophage transition from the M1 to TAM phenotype. To address this question, we next investigated the downstream targets of *RBPJ* to characterise the functional consequences of NOTCH signalling.

***RBPJ* coexpression pinpoints downstream targets of NOTCH signalling and its role in converting macrophages to TAM.** We performed a separate NNet analysis to measure *RBPJ*'s coexpression with 5,011 target genes. By calculating the correlation between *NOTCH2* activity (as shown in Figure 4G) and each target's coexpression with *RBPJ* across individual cells, we identified a set of genes that were highly associated with NOTCH-RBPJ signalling (Figure 4I). According to prior knowledge, key genes triggered by this signalling include *MAF* and *DAB2*, which are important regulators of M2 polarisation (Liu et al., 2020; Marigo et al., 2020), as well as the cell surface receptors CD206 (*MRC1*) and CD209 (*CD209*), which are defining markers for M2-like TAMs (Wang et al., 2024). While Franklin et al. (2014) suggested that *RBPJ*-dependent

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

TAMs are *MRC1*-, our findings indicate that *RBPJ* may induce *MRC1*$^+$ TAMs at a transient state. Functionally, we also noted that NOTCH-RBPJ signalling triggered macrophages to secrete chemoattractants such as *CCL13*, which may recruit immune cells to the TME, thereby modulating the immune landscape (Supplementary Figure S3D).

Taken together, our final case study showcases NNet's ability to capture signalling interactions that precede gene regulation in individual cells. Through this approach, we identify that NOTCH-RBPJ signalling becomes active at an intermediate macrophage state, positioned between the M1 pro-inflammatory and M2-like TAM phenotypes. In this transitional phase, NOTCH-RBPJ may modulate re-education signals from the TME, facilitating the conversion of M1 macrophages into TAM.

Discussion

NeighbourNet (NNet) is a novel and highly scalable framework for constructing cell-specific coexpression networks from scRNA-seq data. Based on PC regression, our approach unravels coexpression within the local neighbourhood of individual cells, offering a fresh perspective for understanding regulatory dynamics beyond traditional clustering-based gene network inference methods. NNet is highly scalable, capable of capturing networks for thousands of cells at low computational cost. A key advantage is that it pairs this scalability with accessibility, offering a comprehensive suite of downstream analyses that enables researchers to fully leverage the granularity of cell-specific networks (CSNs) for meaningful biological discovery. We highlighted the versatility of NNet through three case studies and its utility in facilitating the identification of fine-grained regulatory signals pertinent to specific cell states or transitions.

In the first case study of transcription factor (TF) activity inference, we performed a proof-of-concept analysis on NNet coexpression. By comparing TF activity scores calculated based on TF-target coexpression with those obtained solely from target expression, we demonstrated that coexpression-based scores more accurately reflect TF function and activation status. These findings

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

showed that NNet coexpression carries reliable evidence of gene regulation and highlight that traditional expression-based TF activity inference may yield higher false discovery rates, warranting more cautious interpretation.

In the second case study on early hematopoiesis, we introduced a novel perspective for analysing cellular variation through coexpression analysis. We showed that analyses traditionally performed in gene expression space, such as dimensionality reduction, clustering, and pseudotime analysis can be effectively adapted to coexpression space. A key insight from this study was that analysing coexpression without relying on predefined clusters substantially improves the ability to uncover detailed regulatory signals in scRNA-seq data. The successful identification of the TF module associated with conventional type 1 dendritic cell (cDC1) differentiation exemplified this advantage, as traditional cluster-based approaches tend to obscure features of rare cell populations within broader groupings. While we do not present results that demonstrate novel biological discoveries, many of our findings warrant further investigation. For example, the composition of the meta-TF module and how its factors co-regulate to promote the differentiation of the clinically significant DC subtype deserve closer study.

In the third case study involving small cell lung cancer, we demonstrate NNet's capability to integrate prior regulatory knowledge, ranging from gene regulation to signalling interactions, into the functional characterisation of coexpression networks. Through comprehensive visualisation and quantitative analyses of CSNs, NNet emerges as a pioneering algorithm that resolves fine-grained details of signalling pathways at the individual cell level, addressing a key gap in current computational approaches. We pinpointed a potentially novel role of NOTCH–RBPJ signalling in tumour microenvironment (TME) development through the re-education of M1 pro-inflammatory macrophages. Franklin et al. (2014) reported that this pathway promotes monocyte differentiation into *MRC1*$^-$ tumour-associated macrophages (TAMs). Our results suggest that it may also contribute to the development of *MRC1*$^+$ TAMs. However, we cannot draw a definite conclusion about whether the TAMs we analysed arise from the re-education of existing M1 macrophages or from the primary differentiation of monocytes, simply based on the continuous transition observed between M1 macrophages and TAMs. We also cannot infer that this pathway causally drives the

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

development of *MRC1*$^+$ TAMs, and these hypotheses require experimental validation.

We identified some limitations in our study. First, we did not perform simulation-based benchmarks, which are commonly used to validate new method for inferring global or cluster-specific networks. This omission is primarily due to the substantial technical challenges of simulating realistic CSNs. As a result, our validation used real-world datasets and focused on the basic property of CSNs: their ability to preserve first-level cell type information. Second, our downstream network analysis focused on basic network properties such as node degree, primarily to enhance interpretability. Incorporating additional metrics, such as clustering coefficients, modularity, or network motifs, could further illuminate the regulatory complexity and dynamic characteristics of CSN.

Looking ahead, NNet's framework can be adapted to spatially resolved transcriptomics data, where cell neighbourhoods are defined by spatial proximity rather than gene expression similarity. This extension would enable the study of spatial intercellular communication within tissue-specific microenvironments. Furthermore, by adjusting how data embedding and regression are performed within cell neighbourhoods, NNet can be applied to infer interactions involving molecules beyond transcripts, including ligand–receptor binding (proteomics) and multi-omics relationships. A particularly promising direction is the use of partial least squares (PLS) regression for multi-omics embedding, which could facilitate the integration of diverse data modalities and broaden NNet's applicability across biological contexts.

Material & Methods

All bioinformatic analyses, including the development of the `NeighbourNet` software, were conducted in R version 4.5.0 (R Core Team, 2025).

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1. NeighbourNet regression setting and computational cost: We summarise the number of cells (N), the number of response genes (P), and the number of predictor genes (Q) involved in the NNet analysis for each case study. Each NNet analysis thus generates an ensemble of $N \times P \times Q$ coexpression networks. For NNet analyses in the early hematopoiesis (Case study 2) and lung cancer (Case study 3) case studies, networks were built on a subset of cells (indicated by 'subsampled') to represent the full dataset, further reducing computational burden. The **Computational time** column records the runtime of NNet in seconds, broken down into the stages of local gene variance calculation (LVC), regression, and meta-network construction. LVC calculation is an initial part of the coexpression measurement (Material & Methods) and only needs to be performed once before regression. Adjusting the response genes for subsequent regression steps does not require recalculating local variance. The **Memory usage** column shows the change in memory (in gigabytes) before and after the LVC and regression steps. All the analysis were ran on a RStudio server allocated with four cores of Intel Xeon Gold 6254 @ 3.10GHz CPU. We did not perform parallel computing. Other acronyms: DEGs: differentially expressed genes. PKN: prior knowledge network.

Data	Cells (N)	Responses (P)	Predictors (Q)	Computational time (s)	Memory usage (Gb)
PBMC3k data Case study 1	N = 2,638	P = 288 CollecTRI TFs	Q = 4,116 CollecTRI targets	LVC: 34.56 Regression: 784.52	25.60
Perturb-seq data 1 Case study 1 (Papalexi et al., 2021)	N = 7,411	P = 5 CollecTRI TFs	Q = 5,151 CollecTRI targets	LVC: 155.39 Regression: 18.6	9.35
Perturb-seq data 2 (d7) Supplementary Results (Dixit et al., 2016)	N = 16,506	P = 10 perturbed TFs	Q = 3,227 CollecTRI targets	LVC: 872.87 Regression: 120.48	5.62
Perturb-seq data 2 (d13) Supplementary Results (Dixit et al., 2016)	N = 9,633	P = 10 perturbed TFs	Q = 3.227 CollecTRI targets	LVC: 305.20 Regression: 63.18	3.00
Lin$^-$ early hematopetic cell atlas Case study 2 (Pellin et al., 2019)	N = 1,078 (Subsampled)	P = 805 PKN TFs	Q = 4,600 PKN targets	LVC: 33.68 Regression: 1045.09 Meta-network 482.11	34.50
Small cell lung cancer atlas Case study 3 (Chan et al., 2021)	N = 2,909 (Subsampled)	P = 28 DEGs	Q = 900 PKN TFs	LVC: 87.03 Regression: 66.61 Meta-network 39.11	1.23

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

NeighbourNet: cell-specific coexpression network inference

We develop NeighbourNet (NNet) to infer gene networks from scRNA-seq data. Unlike traditional statistical methods that rely on coarse-grained cell type clusters to measure gene coexpression, NNet directly analyses coexpression within each cell's k-nearest neighbourhood (KNN) in the gene expression space. To improve computational efficiency and mitigate noise in scRNA-seq data, we apply principal component analysis (PCA) to embed and denoise gene expression. Coexpression is then measured between PCs and response genes using regression models, and gene-gene coexpression is subsequently recovered from functional dependencies learned between PCs and response genes. In this section, we outline how we build a coexpression network for each cell (cell-specific network, CSN) and how we refine these networks based on the KNN graph of cells, a cell-by-cell matrix describing KNN relationships in gene expression. Briefly, the procedure consists of the following steps:

1. Build a KNN-graph based on PCA, describing neighbourhood structure of cells.
2. Fit PC regression within each cell's neighbourhood to build CSN.
3. Smooth the networks using the KNN-graph to reduce noise.
4. Calculate the significance of coexpression based on the KNN graph constructed.

In the following sections, we describe the NNet algorithm in more detail with some mathematical formulation. Additional technical details are provided in Supplementary Methods.

Build KNN Graph. We first perform PCA (with centering and scaling) on the $N \times P$ cell-by-gene expression matrix $\mathcal{X}$ to obtain an $N \times R$ score matrix $\mathcal{S}$ representing R principal components (PCs). The choice of R follows Linderman et al. (2022) (Supplementary Methods). Using these PCs, we reconstruct a low-rank approximation (LRA) of the expression data $\hat{\mathcal{X}}$, which serves as the response variable in PC regression.

Next, we construct a weighted KNN graph of cells represented by an $N \times N$ adjacency matrix W. Each cell $n \in \{1, 2, \ldots, N\}$ in the graph is connected to its neighbours $\mathrm{KNN}(n)$ based on geodesic distances in the PCA space, following the approach of Becht et al. (2019) (Supplementary Methods).

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

By default, 30 nearest neighbours are considered. Stronger edge weights correspond to shorter distances between cells. See Supplementary Result and Supplementary Figure S5 for an evaluation of tuning parameters.

PC regression within each cell's neighbourhood. Within each cell n's neighbourhood, we extract the neighbouring cells' LRA and PC scores, denoted by $\hat{\mathbf{X}}$ and $\mathbf{S}$, respectively (suppressing the index n for clarity). For a given gene $p \in \{1, 2, \ldots, P\}$, we treat its reconstructed expression $\mathbf{y} \equiv \hat{\mathbf{X}}_{\cdot p}$ as the response, and fit a local PC regression model:

$$\mathbf{y} \sim f_{np}(\mathbf{S}).$$

where cell $m \in \mathrm{KNN}(n)$ are weighted in the regression according to their connection strength to n, the master cell of the neighbourhood, given by $\mathcal{W}_{nm}$.

Coexpression levels between the response gene p and each predictor gene $q \in \{1, 2, \ldots, P\}$ contributing to PC computing is defined as permutation feature importance (PFI) score. Because direct permutation of thousands of genes within each local neighbourhood is computationally prohibitive, we derive an analytical approximation: PFI is estimated by the dot product between the partial derivatives of $\hat{f}_{np}$ with respect to PCs and each gene q's loading vector. If we use a linear model to fit a local PC regression for response gene p for cell n, with $\hat{\beta}_{npr}$ denoting the fitted regression coefficient of PC r, then the coexpression between an embedded gene q and the response p is defined as

$$\mathrm{CSN}_{npq} = 2\,\mathrm{Var}(\mathbf{X}_{\cdot q}) \left(\sum_r \hat{\beta}_{npr} \frac{\mathcal{V}_{qr}}{\|\mathcal{V}_{q\cdot}\|} \right)^2$$

where $\mathrm{Var}(\mathbf{X}_{\cdot q})$ is the local variance of gene q expression within the neighbourhood, and $\mathcal{V}$ represents the PC loading matrix.

Repeating this procedure across response genes produces one cell-specific coexpression network per cell, represented by an adjacency matrix:

$$\mathrm{CSN}_n \in \mathbb{R}^{P \times P}$$

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Our framework supports PC regression with various regression methods, with ridge regression (penalty parameter = 0.5) used as the default. Full derivations and implementation details are provided in Supplementary Methods.

Smoothing and significance assessment Because local regression is performed within relatively small neighbourhoods, coexpression estimates may be sensitive to sampling variability. To reduce instability, CSNs are smoothed on the KNN graph of the cells using a random walk operator derived from W, which propagates information across neighbouring cells while preserving local structure.

Statistical significance of each coexpression edge is then assessed using a neighbourhood-agnostic null model. Specifically, we construct a null distribution by recomputing coexpression after shuffling the KNN graph structure, thereby destroying true local dependencies while preserving marginal distributions. This generates a background distribution of smoothed coexpression values under the absence of structured neighbourhood effects. For each edge, the observed coexpression is compared against this null distribution, yielding a Wald-type test statistic and a probabilistic significance score between 0 and 1. Users may prune CSNs by applying a significance threshold prior to downstream analyses.

The algorithmic details are provided in Supplementary Methods. An assessment of the proposed significance score is performed on an independent dataset from Tian et al. (2019) (Supplementary Result; Supplementary Figure S6).

Restrict NNet analysis on subsampled cells to reduce computational demand. To further reduce the computational demands of NNet, both in terms of time and memory usage, and to make it feasible for use on personal computers, we infer CSNs on a representative subset of cells from the entire dataset. To select representative cells, we apply k-means clustering to the PCs, setting the number of clusters to match the desired number of representative cells. For each cluster, we select the cell closest to the cluster centroid as the representative.

NNet runs PC regression and build CSNs on the selected representative cells, leveraging their neighbouring cells from the complete dataset. This yields a smaller network ensemble on which downstream analysis can then be applied. For details on how we perform data smoothing on this

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

network ensemble, refer to Supplementary Methods.

NNet downstream analysis

Beyond coexpression, NNet supports two major downstream analyses (Figure 1D). We first provide an overview of the two analyses without diving into technical details.

The first is the use of non-negative matrix factorisation (NMF) to derive soft clusters of cells or genes from the ensemble of CSNs. Soft cell clusters are represented by cell weights, which are used to construct meta-networks via weighted averaging of CSNs, thereby capturing overarching coexpression patterns across groups of cells. Similarly, genes can be clustered and aggregated based on their weights that represent gene modules, creating meta-genes and their coexpression profiles across cells. We optimised the NMF algorithm to make it scales efficiently on the large ensemble we generate.

Second, we integrate prior knowledge to annotate coexpression edges. We construct integrated prior knowledge networks (PKNs) following the NicheNet framework (Browaeys et al., 2020; Müller-Dott et al., 2023), combining curated regulatory and signalling interactions. Coexpression edges supported by prior knowledge are assigned regulatory directionality to generate context-specific gene regulatory networks (GRNs). Building on these networks, we infer upstream signalling pathways (USPs) that connect receptor signal transduction to TF-mediated target regulation. NNet further provides visualisation tools to explore inferred gene regulation and signalling interaction.

Embed network ensembles. A three-dimensional network ensemble can be embedded into two-dimensional meta-networks using NMF.

Specifically, each CSN is vectorised into a column of a $P^2 \times N$ 'tall-and-skinny' matrix $\mathcal{A}^{\text{cell}}$ where row correspond to edges and columns correspond to individual cells. NMF decomposes this matrix as

$$\mathcal{A}^{\text{cell}} = \mathcal{F}\mathcal{H}^T$$

where $\mathcal{F}$ is a $P^2 \times N'$ factor matrix representing N' vectorised meta-networks, and $\mathcal{H}$ is a $N \times N'$

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

factor matrix representing N' soft clusters.

To address the computational challenges of applying NMF to large matrices such as $\mathcal{A}^{\text{cell}}$, we employ a computationally efficient variant of NMF based on non-negative principal component analysis (nPCA) (Sigg and Buhmann, 2008), a PCA formulation that enforces non-negative loadings. The computational trick of (Benson et al., 2014) was incorporated into the nPCA algorithm, allowing us to perform factorisation on a reduced representation of $\mathcal{A}^{\text{cell}}$ without explicitly constructing the full matrix, thereby saving both computation time and memory.

A similar nPCA approach is used for meta-gene (meta-TF) analysis (Case study 2). Instead of retaining the cell dimension, the network ensemble is converted into an $NP \times P$ tall matrix $\mathcal{A}^{\text{response}}$ that preserves the response (or predictor) gene dimension. Applying nPCA to this matrix clusters and embeds responses, generating a factor matrix that represent coexpression profile between meta-response and predictors across the cells. Refer to Supplementary Methods for implementation details.

Construct integrated prior knowledge networks. We performed the NicheNet integration of PKN using the R package `Omnipath`, which not only provides a collection of high-quality databases but also offers a framework for users to choose their own databases with prior knowledge confidence levels (Müller-Dott et al., 2023). See Supplementary Methods for details on our database selection process. We obtained two integrated PKNs as directed graphs:

- **TF-target gene regulatory network**: Comprising 1,176 TFs, 6,568 targets, and 41,206 gene regulation interactions.
- **Signalling interaction network**: Incorporating 9,867 genes with 79,242 signalling interactions.

These PKNs were pre-computed and edge-annotated to depict whether they consistently indicate stimulation or inhibition, as aligned in the source databases. Additionally, by executing a personalised PageRank algorithm on the signalling PKN, we pre-computed a matrix reflecting the prior regulatory potential of receptors on TFs (Supplementary Methods) (Browaeys et al., 2020; Page, 1999):

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

- **Regulatory potential matrix**: Comprising the regulatory potential of 415 receptors on 905 TFs, aiding in USP inference.

This matrix will be used to facilitate the USP inference.

Receptor prioritisation USP inference aims to reconstruct potential intracellular signalling pathways linking receptor activation to TF-mediated regulation of downstream target genes, as reflected by TF–target coexpression in the CSN. USP inference proceeds in two stages. First, we integrate the signalling interaction PKN with CSNs to prioritise receptors based on their potential activity in regulating downstream targets through TFs. Second, for prioritised receptors, we reconstruct the shortest signalling paths linking receptors to TFs using the PKN.

Receptor prioritisation follows NicheNet's ligand activity framework, adapted to our receptor–TF–target setting. Specifically, we compute a receptor–target activity matrix of a cell by taking the dot product between the receptor–TF regulatory potential matrix and the adjacency matrix of a pruned CSN (Figure 4A2). By summing a receptor's per-target activity values over all targets, we compute an overall receptor activity score that reflects the strength of the receptor's regulatory impact on the entire target gene set. Receptors with high activity scores are considered key upstream regulators and are thus prioritised for further exploration. Implementation details are provided in Supplementary Methods.

Intuitively, the receptor activity measures the effectiveness of a two-stage signal transduction: from receptor to TF through the PKN, and from TF to target through coexpression edges in the CSN. Its reliability, therefore, depends critically on the confidence in these edges as proxies for regulatory interactions. To control this, we apply pruning to CSNs prior to downstream analysis at three levels of stringency:

1. **No pruning**: retains all coexpression edges.
2. **Statistical pruning**: removes edges that fail significance testing.
3. **High-confidence pruning (default)**: retains only edges that are both statistically significant and supported by prior regulatory knowledge.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

An edge is considered supported if the corresponding TF can transmit regulatory influence to the target within a defined number of steps (two steps by default) in the gene regulation PKN.

Reconstruct upstream signalling pathways (USPs). For each prioritised receptor in each CSN, we first identify TFs that plausibly mediate the signal transduction to targets. Candidate TFs are selected based on having strong node degree in the pruned CSN, and receiving high regulatory potential from the receptor (Supplementary Methods).

For each selected receptor–TF pair, we then use the `igraph::shortest_paths` function (R package `igraph`, v2.0.1.1) to identify minimal signalling paths within the signalling interaction PKN, which represents the most direct series of known molecular interactions linking receptor activation to TF modulation. Finally, combining these receptor–TF paths with coexpression-supported TF–target interactions yields receptor–TF–target triplets, representing putative intracellular signalling pathways inferred for each cell.

Visualise prior knowledge enriched networks. We visualise prior-knowledge-enriched TF-target coexpression networks using a hierarchical structure, as shown in Figure 4E.

1. **Target genes**: positioned at the inner most layer.
2. **TFs**: encircle targets, and are grouped by expression-based clustering (node color), assigned significance probabilities (pie charts: % fill = likelihood of coexpression with targets).
3. **USPs**: The outermost layer lays the most active receptors, each connected to a TF that likely mediates the receptor's influence on the targets via USP inference. Receptors are coloured to match TFs, with the percentage of color filling represents the expression level. When a receptor and a TF lack a direct link, USP inference puts an extra layer between them shows their shortest signalling path.

Edges indicate evidence of interaction: none (no edge), significant coexpression only (dashed), or significant coexpression supported by prior evidence (solid, with arrowhead for activation or barhead for repression).

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Central gene selection. In networks containing thousands of genes, it is difficult to visualise the entire structure, as the dense overlap of edges becomes messy and obscures meaningful information. Therefore, it is often more informative to visualise subnetworks consisting of a subset of central genes (for example, in Figure 4E, only central TFs are shown).

We identify central genes by evaluating eigenvector centrality in meta-networks. We apply SVD to each meta-network's response-by-predictor adjacency matrix. The absolute values of the left and right singular vectors correspond to the centrality of responses and predictors, respectively. Genes with the highest centrality are selected, and the union of top-ranked genes across multiple meta-networks defines the final set of central genes.

Data for NNet analysis

We provide an overview of the public scRNA-seq datasets utilised in our case study and detail the preprocessing steps performed prior to NNet analysis. All data were preprocessed and analysed based on a `Seurat` object in R (`Seurat` package version: 5.0.3). Unless specifically noted, NNet was configured using the default tuning parameters as described in the methodology section below.

The 10x PBMC3k data We obtained a widely recognised dataset of human peripheral blood mononuclear cells (PBMC), commonly used for benchmarking cell clustering methods, from the Seurat PBMC3k tutorial at `https://satijalab.org/seurat/articles/pbmc3k_tutorial` (Last retrieved: Oct 27th, 2024) (Butler et al., 2018). Following the code outlined in the vignette, we preprocessed the data, performed dimension reduction using PCA and UMAP, and conducted clustering. The resulting data is log-normalised, containing 2,638 cells and 13,714 genes.

We treated TFs as responses and target genes as predictors in our NNet analysis. Since our TF activity inference relies on gene regulation data from the CollecTRI database (Müller-Dott et al., 2023), we extracted the relevant TF and target gene sets from CollecTRI. Among the 13,714 genes in the dataset, 731 TFs and 4,116 target genes were identified. To ensure meaningful inference of TF activity, we further filtered the 731 TFs to include only those with at least 10 associated targets according to CollecTRI prior knowledge. This resulted in 288 TFs for NNet analysis on all the cells, maintaining 4,116 target genes as predictors.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

The Perturb-seq data The data were originally generated by Papalexi et al. (2021), who performed CRISPR screening on the THP-1 monocyte cell line to investigate molecular networks regulating PD-L1 expression in acute leukemia, a potential mechanism of T cell inhibition. The original study perturbed 26 genes that are correlated with PD-L1 expression. We used a cleaned version of this dataset obtained from the *PerturbSeq.db* database (He et al., 2025). The processed dataset contains 18,257 genes across 7,411 cells, each subjected to one of 10 filtered valid perturbations or left unperturbed as negative controls. Among these 10 genes, 5 are validated TFs: *BRD4*, *IRF1*, *SMAD4*, *STAT1*, and *STAT2*. We did not perform further quality control filtering on genes or cells.

NNet analysis was performed on all cells, using the 5 perturbed TFs as responses. Following the same procedure as with the PBMC dataset, we selected 5,151 CollecTRI target genes as predictors.

The early hematopoiesis atlas The data generated by Pellin et al. (2019) profiles an early human hematopoietic landscape of bone marrow mononuclear cells that lack mature lineage markers (Lin^-). We obtained the raw data from the GEO accession GSE117498 and downloaded the four Lin^- samples. These samples were merged into a combined data comprising 24,720 genes and 15,397 cells. Without performing additional quality control to filter out genes or cells, we applied the Seurat pipeline to process the data. We annotated cell clusters according to the marker genes provided by Pellin et al. (2019).

The NNet analysis was conducted on a subset of 1,078 cells, using 805 TFs as responses and 4,600 targets as predictors. These TF and target gene sets were collected from our PKN with the criterion that they should be expressed by at least 20 cells.

The small cell lung cancer atlas The dataset was generated by Chan et al. (2021), who aimed to profile the heterogeneity of small cell lung cancer (SCLC) and its associated microenvironment across different lung cancer subtypes, with a specific focus on identifying and characterising macrophage subpopulations linked to the presence of a metastatic SCLC subtype. Additionally, they collected and sequenced specimens from lung adenocarcinoma (LUAD) and normal lung tissues for comparative analysis. We downloaded the processed data from the CELLxGENE data portal,

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

which comprises cells pooled from 42 donors. We retained 19,558 genes that have HUGO symbols for easier interpretation and did not perform additional quality control to filter out cells from the processed data. Utilising the cell type annotations provided by CELLxGENE, we extracted 9,251 macrophages, on which we then applied the Seurat pipeline to process the data.

We conducted two separate NNet analyses on the same set of 2,909 macrophage that were subsampled: one aimed at identifying signalling interactions associated with pro-tumorigenic macrophage development, and another focused on determining the downstream targets of NOTCH-RBPJ signalling. For the first analysis, response genes were identified through a differential expression (DE) analysis comparing macrophages from tumour and normal samples. From the top 50 differentially expressed genes (DEGs) of tumour macrophages, we selected 28 genes with known TF regulators according to our PKN to serve as responses for NNet analysis. In the second analysis, *RBPJ* was set as the NNet response. When selecting predictors, we only considered genes that were expressed by at least 20 macrophages. As a result, 900 TFs and 5,011 targets (not specific to *RBPJ*) in the PKN were selected as predictors for the first and second analyses respectively.

AUCell and decoupleR analysis for TF activity inference

The R package `decoupleR` (Version 2.9.7) provides an ensemble of computational approaches for TF activity inference based on gene expression data (Badia-i Mompel et al., 2022). This ensemble includes the AUCell method and the package's new approach that relies on univariate linear regression, referred to as the decoupleR method in our paper (Aibar et al., 2017). Both AUCell and the decoupleR method infer TF activity by evaluating how well a cell's gene expression can predict the known target gene set of a TF. AUCell calculates the area under the curve (AUC) score for the sorted expression of a cell in distinguishing the target gene set. In contrast, the decoupleR method performs a regression between gene expression and a dummy vector representing the target gene set, calculating TF activity as the t-value of the regression. Both functions take a gene expression matrix and a prior knowledge GRN as inputs, iterate AUC and t-value calculations on each cell for each TF, eventually generating a TF-by-cell matrix of TF activity. The prior knowledge GRN we used was generated by `decoupleR::get_collectri` based on the CollecTRI database, which was

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

in a specific data format required by the decoupleR functions.

TF activity inference based on TF-target coexpression was also performed using the decoupleR package. The only difference was that, for each TF, we measured its coexpression with targets across cells, generating a TF-specific coexpression profile on which activity inference was then performed.

Evaluate the effectiveness of TF activity in distinguishing perturbed cells using area under the curve

For the Perturb-Seq data, we used AUCell and decoupleR to measure the activity of 5 TFs that were perturbed. For each of the 5 TFs, we compared the activity scores between two groups of cells: those perturbed exclusively on that TF and those that were not perturbed on any TF (controls). The area under the curve (AUC) was calculated to assess how well the activity score of the perturbed TF could distinguish the unperturbed cells. AUC evaluations for the smaller batch clusters are provided in Supplementary Figure S1. AUC was calculated using the R package `pROC` (Version 1.18.0) with the function `pROC::roc`.

Embed and cluster PBMC3K data by TF activity

On the PBMC data, we calculated TF activity for 288 TFs across 2,638 cells based on different activity inference methods (AUCell and decoupleR) applied to different measurements (expression and coexpression). The resulting TF activity matrices were then embedded using PCA (centred and scaled) followed by UMAP (using 10 PCs).

We employed the adjusted Rand index (ARI) and median silhouette index (MSI) to quantitatively assess how well the activity scores of 288 TFs recapitulate the cell clusters we learnt from the expression of the 2000 most variable features (Rand, 1971; Rousseeuw, 1987). MSI was calculated based on the distances between cells in the 10 PC space of activity scores, evaluating whether cells within the same gene expression cluster are also grouped together according to their TF activity profiles. ARI is a metric that evaluates the agreement between two different clustering schemes. We utilised ARI to compare clusters generated from activity scores with those derived

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

from gene expression data. For the clustering based on activity scores, we employed Seurat's clustering pipeline to generate exactly 9 clusters, matching the number of clusters obtained from the expression data.

Over-representation analysis on meta-TF weighting vectors

We performed an over-representation analysis on the NMF weighting vector of meta-TF 7, the first meta-TF that showed strong connections with target genes in the mononuclear-phagocyte lineage. Using the R package `clusterProfiler` (version 4.11.1), we identified over-represented Gene Ontology (GO) terms in biological processes among the top 50 weighted TFs of meta-TF 7 (Yu et al., 2012). A false discovery rate threshold of 0.05 was set to determine significant GO terms, and the background gene set for the analysis consisted of the 805 TFs used in the NMF embedding.

The resulting GO terms were ranked according to their pi-values, calculated as the product of their fold changes and their negative $\log_{10}$ p-values (Xiao et al., 2014). Highly redundant GO terms were trimmed by the `clusterProfiler::simplify` function. Finally, the top five non-redundant GO terms with the highest pi-values were illustrated.

Diffusion pseudotime inference

Using the R package `destiny` (Version 3.10.0), we performed diffusion pseudotime (DPT) inference on the early hematopoiesis atlas to reconstruct the temporal ordering of cells during differentiation (Haghverdi et al., 2015, 2016). Briefly, DPT for each cell was derived from its distance to a user-selected root cell, representing the earliest differentiation stage, within a diffusion map (DM) embedding of the data. To create the DM, we applied `destiny::DiffusionMap` to the 10 PCs of the 2,000 most variable genes computed previously. The resulting DM served as the input for `destiny::DPT` to infer DPT. We selected the root cell as the one with the highest *CD34* expression, a marker for hematopoietic stem cells.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Differential expression analysis on macrophages within lung tumour

Macrophages from SCLC and LUAD samples were grouped together for comparison with normal macrophages, as differences between cancer subtypes were not of interest. The DE analysis was performed using the Seurat function `Seurat::FindMarkers`, selecting MAST as the DE method and retaining DEGs that were expressed in at least 10 percent of the tumour macrophages (Finak et al., 2015). The top 50 DEGs of the tumour macrophages were then selected according to their pi-values (Xiao et al., 2014).

Competing interest statement

The authors declare that they have no competing interests.

Code availability

All code required to reproduce the analyses and the NeighbourNet R package are available at GitHub (`https://github.com/meiosis97/NeighbourNet`) and as Supplemental Code.

Acknowledgements

Author Contributions

YD conceived the study, developed the methods, performed the formal analyses, and wrote and edited the manuscript. JM contributed to the study conception and manuscript editing. JC and KALC edited the manuscript.

Funding

JM and KALC were supported in part by the National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2025648). YD was supported by the Melbourne Research Scholarship.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). Scenic: single-cell regulatory network inference and clustering. Nature methods, 14(11):1083–1086.

Anastasiou, G., Gialeraki, A., Merkouri, E., Politou, M., and Travlou, A. (2012). Thrombomodulin as a regulator of the anticoagulant pathway: implication in the development of thrombosis. Blood Coagulation & Fibrinolysis, 23(1):1–10.

Badia-i Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., et al. (2022). decoupler: ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances, 2(1):vbac016.

Badia-i Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., and Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. Nature Reviews Genetics, 24(11):739–754.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology, 37(1):38–44.

Benson, A. R., Lee, J. D., Rajwa, B., and Gleich, D. F. (2014). Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices. Advances in neural information processing systems, 27.

Bill, R., Wirapati, P., Messemaker, M., Roh, W., Zitti, B., Duval, F., Kiss, M., Park, J. C., Saal, T. M., Hoelzl, J., et al. (2023). Cxcl9: Spp1 macrophage polarity identifies a network of cellular programs that control human cancers. Science, 381(6657):515–524.

Böttcher, J. P., Bonavita, E., Chakravarty, P., Blees, H., Cabeza-Cabrerizo, M., Sammicheli, S., Rogers, N. C., Sahai, E., Zelenay, S., and e Sousa, C. R. (2018). Nk cells stimulate recruitment of cdc1 into the tumor microenvironment promoting cancer immune control. Cell, 172(5):1022–1037.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Bresnick, E. H., Hewitt, K. J., Mehta, C., Keles, S., Paulson, R. F., and Johnson, K. D. (2018). Mechanisms of erythrocyte development and regeneration: implications for regenerative medicine and beyond. Development, 145(1):dev151423.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). Nichenet: modeling intercellular communication by linking ligands to target genes. Nature methods, 17(2):159–162.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, 36(5):411–420.

Chan, J. M., Quintanal-Villalonga, A., Gao, V. R., Xie, Y., Allaj, V., Chaudhary, O., Masilionis, I., Egger, J., Chow, A., Walle, T., et al. (2021). Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. Cancer cell, 39(11):1479–1496.

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems, 5(3):251–267.

Chen, S., Saeed, A. F., Liu, Q., Jiang, Q., Xu, H., Xiao, G. G., Rao, L., and Duo, Y. (2023). Macrophages in immunoregulation and therapeutics. Signal Transduction and Targeted Therapy, 8(1):207.

Collin, M. and Bigley, V. (2018). Human dendritic cell subsets: an update. Immunology, 154(1):3–20.

Combes, T. W., Orsenigo, F., Stewart, A., Mendis, A. J. R., Dunn-Walters, D., Gordon, S., and Martinez, F. O. (2021). Csf1r defines the mononuclear phagocyte system lineage in human blood in health and covid-19. Immunotherapy Advances, 1(1):ltab003.

Dai, H., Li, L., Zeng, T., and Chen, L. (2019). Cell-specific network constructed by single-cell rna sequencing data. Nucleic acids research, 47(11):e62–e62.

Delgado, M. D. and León, J. (2010). Myc roles in hematopoiesis and leukemia. Genes & cancer, 1(6):605–616.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. cell, 167(7):1853–1866.

Dwyer, A. R., Greenland, E. L., and Pixley, F. J. (2017). Promotion of tumor invasion by tumor-associated macrophages: the role of csf-1-activated phosphatidylinositol 3 kinase and src family kinase motility signaling. Cancers, 9(6):68.

Elahi, Z., Angel, P. W., Butcher, S. K., Rajab, N., Choi, J., Deng, Y., Mintern, J. D., Radford, K., and Wells, C. A. (2022). The human dendritic cell atlas: an integrated transcriptional tool to study human dendritic cell biology. The Journal of Immunology, 209(12):2352–2361.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology, 16:1–13.

Foldi, J., Shang, Y., Zhao, B., Ivashkiv, L. B., and Hu, X. (2016). Rbp-j is required for m2 macrophage polarization in response to chitin and mediates expression of a subset of m2 genes. Protein & cell, 7(3):201–209.

Franklin, R. A., Liao, W., Sarkar, A., Kim, M. V., Bivona, M. R., Liu, K., Pamer, E. G., and Li, M. O. (2014). The cellular and molecular origin of tumor-associated macrophages. Science, 344(6186):921–925.

Friedrich, T., Ferrante, F., Pioger, L., Nist, A., Stiewe, T., Andrau, J.-C., Bartkuhn, M., Giaimo, B. D., and Borggrefe, T. (2022). Notch-dependent and-independent functions of transcription factor rbpj. Nucleic Acids Research, 50(14):7925–7937.

Hagemann, T., Biswas, S. K., Lawrence, T., Sica, A., and Lewis, C. E. (2009). Regulation of macrophage function in tumors: the multifaceted role of nf-κb. Blood, The Journal of the American Society of Hematology, 113(14):3139–3146.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics, 31(18):2989–2998.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature methods, 13(10):845–848.

He, T., Yang, X., Tong, Y., Liu, X., Wei, Y., Wang, W., Yuan, J., Wang, Y., and Yang, Y. (2025). Perturbseq. db: An integrated repository for comprehensive analysis of single-cell perturbation data. Journal of Molecular Biology, page 169209.

Hildner, K., Edelson, B. T., Purtha, W. E., Diamond, M., Matsushita, H., Kohyama, M., Calderon, B., Schraml, B. U., Unanue, E. R., Diamond, M. S., et al. (2008). Batf3 deficiency reveals a critical role for cd8α+ dendritic cells in cytotoxic t cell immunity. Science, 322(5904):1097–1100.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PloS one, 5(9):e12776.

Jackson, J. T., Hu, Y., Liu, R., Masson, F., d'Amico, A., Carotta, S., Xin, A., Camilleri, M. J., Mount, A. M., Kallies, A., et al. (2011). Id2 expression delineates differential checkpoints in the genetic program of cd8α+ and cd103+ dendritic cell lineages. The EMBO journal, 30(13):2690–2704.

Jayapal, S. R., Lee, K. L., Ji, P., Kaldis, P., Lim, B., and Lodish, H. F. (2010). Down-regulation of myc is essential for terminal erythroid maturation. Journal of Biological Chemistry, 285(51):40252–40265.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. Nature, 614(7949):742–751.

Kassouf, M. T., Hughes, J. R., Taylor, S., McGowan, S. J., Soneji, S., Green, A. L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of tal1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome research, 20(8):1064–1083.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Knox, J. J., Myles, A., and Cancro, M. P. (2019). T-bet+ memory b cells: generation, function, and fate. Immunological reviews, 288(1):149–160.

Kurotaki, D., Yamamoto, M., Nishiyama, A., Uno, K., Ban, T., Ichino, M., Sasaki, H., Matsunaga, S., Yoshinari, M., Ryo, A., et al. (2014). Irf8 inhibits c/ebpα activity to restrain mononuclear phagocyte progenitors from differentiating into neutrophils. Nature communications, 5(1):4978.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics, 9(1):1–13.

Liang, X., Wang, Z., Dai, Z., Zhang, H., Zhang, J., Luo, P., Liu, Z., Liu, Z., Yang, K., Cheng, Q., et al. (2023). Glioblastoma glycolytic signature predicts unfavorable prognosis, immunological heterogeneity, and eno1 promotes microglia m2 polarization and cancer cell malignancy. Cancer Gene Therapy, 30(3):481–496.

Linderman, G. C., Zhao, J., Roulis, M., Bielecki, P., Flavell, R. A., Nadler, B., and Kluger, Y. (2022). Zero-preserving imputation of single-cell rna-seq data. Nature communications, 13(1):192.

Liu, M., Tong, Z., Ding, C., Luo, F., Wu, S., Wu, C., Albeituni, S., He, L., Hu, X., Tieri, D., et al. (2020). Transcription factor c-maf is a checkpoint that programs macrophages in lung cancer. The Journal of clinical investigation, 130(4):2081–2096.

Marigo, I., Trovato, R., Hofer, F., Ingangi, V., Desantis, G., Leone, K., De Sanctis, F., Ugel, S., Canè, S., Simonelli, A., et al. (2020). Disabled homolog 2 controls prometastatic activity of tumor-associated macrophages. Cancer discovery, 10(11):1758–1773.

Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., and Swarup, V. (2023). hdwgcna identifies co-expression networks in high-dimensional transcriptomics data. Cell reports methods, 3(6).

Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R. O., Badia-i Mompel, P., Fallegger, R., Türei, D., Lægreid, A., and Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. Nucleic acids research, 51(20):10934–10949.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Nguyen, H., Tran, D., Tran, B., Pehlivan, B., and Nguyen, T. (2021). A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. Briefings in bioinformatics, 22(3):bbaa190.

Page, L. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Technical Report.

Palaga, T., Wongchana, W., and Kueanjinda, P. (2018). Notch signaling in macrophages in the context of cancer immunity. Frontiers in immunology, 9:652.

Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck III, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., et al. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nature genetics, 53(3):322–331.

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S. L., Montepeloso, A., Weinberg, O. K., Biffi, A., Klein, A. M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nature communications, 10(1):2395.

Plebanek, M. P., Bhaumik, D., Bryce, P. J., and Thaxton, C. S. (2018). Scavenger receptor type b1 and lipoprotein nanoparticle inhibit myeloid-derived suppressor cells. Molecular cancer therapeutics, 17(3):686–697.

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature methods, 17(2):147–154.

Qi, J., Sun, H., Zhang, Y., Wang, Z., Xun, Z., Li, Z., Ding, X., Bao, R., Hong, L., Jia, W., et al. (2022). Single-cell and spatial analysis reveal interaction of fap+ fibroblasts and spp1+ macrophages in colorectal cancer. Nature communications, 13(1):1742.

R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846–850.

Rosa, F. F., Pires, C. F., Kurochkin, I., Halitzki, E., Zahan, T., Arh, N., Zimmermannová, O., Ferreira, A. G., Li, H., Karlsson, S., et al. (2022). Single-cell transcriptional profiling informs efficient reprogramming of human somatic cells to cross-presenting dendritic cells. Science immunology, 7(69):eabg5539.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65.

Satpathy, A. T., Wu, X., Albring, J. C., and Murphy, K. M. (2012). Re (de) fining the dendritic cell lineage. Nature immunology, 13(12):1145–1154.

Sigg, C. D. and Buhmann, J. M. (2008). Expectation-maximization for sparse and non-negative pca. In Proceedings of the 25th international conference on Machine learning, pages 960–967.

Suzuki, M., Kobayashi-Osaki, M., Tsutsumi, S., Pan, X., Ohmori, S., Takai, J., Moriguchi, T., Ohneda, O., Ohneda, K., Shimizu, R., et al. (2013). Gata factor switching from gata 2 to gata 1 contributes to erythroid differentiation. Genes to Cells, 18(11):921–933.

Szabo, S. J., Kim, S. T., Costa, G. L., Zhang, X., Fathman, C. G., and Glimcher, L. H. (2000). A novel transcription factor, t-bet, directs th1 lineage commitment. Cell, 100(6):655–669.

Tallack, M. R., Whitington, T., Yuen, W. S., Wainwright, E. N., Keys, J. R., Gardiner, B. B., Nourbakhsh, E., Cloonan, N., Grimmond, S. M., Bailey, T. L., et al. (2010). A global role for klf1 in erythropoiesis revealed by chip-seq in primary erythroid cells. Genome research, 20(8):1052–1063.

Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A., Jabbari, J. S., et al. (2019). Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. Nature methods, 16(6):479–487.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Tussiwand, R. and Gautier, E. L. (2015). Transcriptional regulation of mononuclear phagocyte development. Frontiers in immunology, 6:533.

Wang, L., Trasanidis, N., Wu, T., Dong, G., Hu, M., Bauer, D. E., and Pinello, L. (2023). Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. Nature Methods, 20(9):1368–1378.

Wang, P., Wang, Z., and Liu, J. (2020). Role of hdacs in normal and malignant hematopoiesis. Molecular cancer, 19:1–21.

Wang, S., Wang, J., Chen, Z., Luo, J., Guo, W., Sun, L., and Lin, L. (2024). Targeting m2-like tumor-associated macrophages is a potential therapeutic approach to overcome antitumor drug resistance. NPJ Precision Oncology, 8(1):31.

Wang, X., Choi, D., and Roeder, K. (2021). Constructing local cell-specific networks from single-cell data. Proceedings of the National Academy of Sciences, 118(51):e2113178118.

Xiao, Y., Hsiao, T.-H., Suresh, U., Chen, H.-I. H., Wu, X., Wolf, S. E., and Chen, Y. (2014). A novel significance score for gene selection and ranking. Bioinformatics, 30(6):801–807.

Xu, H., Zhu, J., Smith, S., Foldi, J., Zhao, B., Chung, A. Y., Outtz, H., Kitajewski, J., Shi, C., Weber, S., et al. (2012). Notch–rbp-j signaling regulates the transcription factor irf8 to promote inflammatory macrophage polarization. Nature immunology, 13(7):642–650.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. Omics: a journal of integrative biology, 16(5):284–287.

Yu, H., Lin, L., Zhang, Z., Zhang, H., and Hu, H. (2020). Targeting nf-κb pathway for the therapy of diseases: mechanism and clinical study. Signal transduction and targeted therapy, 5(1):209.

Zhang, S., Pyne, S., Pietrzak, S., Halberg, S., McCalla, S. G., Siahpirani, A. F., Sridharan, R., and Roy, S. (2023). Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. Nature Communications, 14(1):3064.

Downloaded from genome.cshlp.org on June 22, 2026 . Published by Cold Spring Harbor Laboratory Press

Zhang, S. Y. and Stumpf, M. P. (2023). Learning cell-specific networks from dynamics and geometry of single cells. bioRxiv, pages 2023–01.

Zhang, Z., Han, J., Song, L., and Zhang, X. (2022). Cespgrn: Inferring cell-specific gene regulatory networks from single cell multi-omics and spatial data. bioRxiv, pages 2022–03.